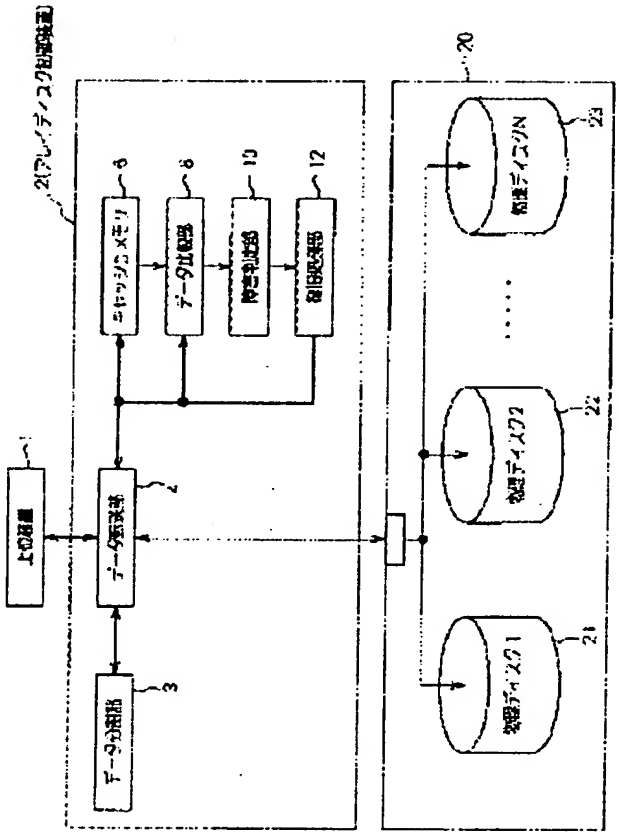
METHOD AND DEVICE FOR CONTROLLING ARRAY DISK

Publication number: JP2001142650
Publication date: 2001-05-25
Inventor: ASANO AKIHIRO
Applicant: NIPPON ELECTRIC CO
Classification:
- international: ***G06F3/06; G11B19/02; G11B20/18;*** **G06F3/06; G11B19/02; G11B20/18;** (IPC1-7): G06F3/06; G11B19/02; G11B20/18
- european:
Application number: JP19990327642 19991118
Priority number(s): JP19990327642 19991118

Abstract of **JP2001142650**

PROBLEM TO BE SOLVED: To find the occurrence of illegal data and a disk failure at an early stage and to reliably store even data that are not read for a long time. SOLUTION: This device is provided with a cache memory 6, a data dividing part 3 which divides data transmitted from a host device 1 in accordance with the number of physical disks and also generates the parity data of the divided division data, a data transferring part 4 which respectively stores the division data and the parity data into physical disks, a data comparing part 8 which reads the data from the physical disk after storing the division data and the parity data in each physical disk and compares the data with corresponding data stored in the cache memory, and a failure deciding part 10 which decides that a failure takes place in the physical disk storing an uncoincidental part when any uncoincidental part exists in the part 8.



(19)日本国特許庁（ＪＰ）　(12) **公開特許公報（Ａ）**　(11)特許出願公開番号
特開2001－142650
（P2001－142650A）
(43)公開日　平成13年5月25日(2001.5.25)

(51)Int.Cl.⁷	識別記号	ＦＩ	テーマコード（参考）
Ｇ０６Ｆ　3/06	３０１	Ｇ０６Ｆ　3/06	３０１Ｒ　５Ｂ０６５
			３０１Ｓ　５Ｄ０６６
	３０５		３０５Ｋ
	５４０		５４０
Ｇ１１Ｂ　19/02	５０１	Ｇ１１Ｂ　19/02	５０１Ｆ

審査請求　有　請求項の数５　ＯＬ　（全７頁）　最終頁に続く

(21)出願番号　特願平11－327642

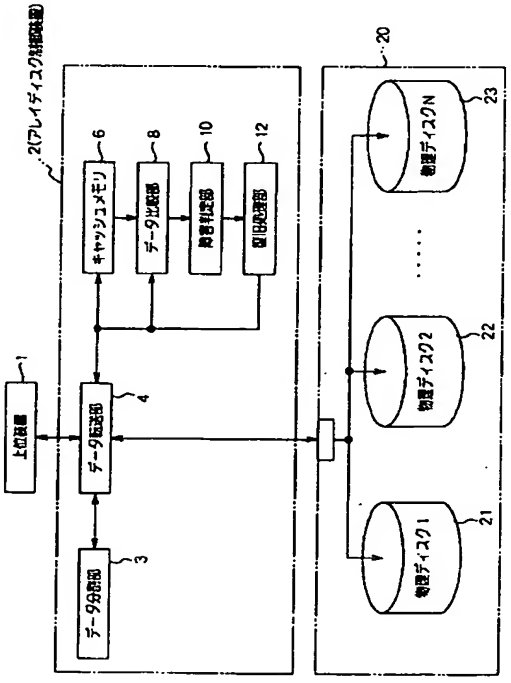
(22)出願日　平成11年11月18日(1999.11.18)

(71)出願人　000004237
日本電気株式会社
東京都港区芝五丁目７番１号
(72)発明者　浅野　聡宏
東京都港区芝五丁目７番１号　日本電気株式会社内
(74)代理人　100079164
弁理士　高橋　勇

Ｆターム(参考)　5B065 BA01 CA13 CA30 CC08 CE12 EA04 EA25
5D066 BA02 BA07 BA08

(54)【発明の名称】　アレイディスク制御方法及び装置

(57)【要約】
【課題】　不正データの発生及びディスク障害を早期に発見して長期間読み出されないデータであっても確実に保存すること。
【解決手段】　キャッシュメモリ６と、上位装置１から送信されたデータを物理ディスクの台数に応じて分割すると共に当該分割した分割データのパリティデータを生成するデータ分割部３と、この分割データ及びパリティデータを各物理ディスクにそれぞれ格納するデータ転送部４と、分割データ及びパリティデータとを各物理ディスクに格納した後、当該各物理ディスクから当該データを読み出してキャッシュメモリに格納された対応するデータと比較するデータ比較部８と、このデータ比較部８で一致しない部分がある場合には当該一致しない部分を記憶した物理ディスクに障害が発生したと判定する障害判定部１０とを備えた。



【特許請求の範囲】

【請求項１】　論理ディスクを構成する複数台の物理ディスクと、上位装置から送信されたデータを一時的に記憶するキャッシュメモリと、前記上位装置から送信されたデータを前記物理ディスクの台数に応じて分割すると共に当該分割した分割データのパリティデータを生成するデータ分割部と、このデータ分割部によって生成された分割データ及びパリティデータを前記各物理ディスクにそれぞれ格納するデータ転送部とを備えたディスクアレイ装置を使用して上位装置から送信されたデータを格納するディスクアレイ制御方法であって、

前記分割データ及びパリティデータの前記各物理ディスクへの格納処理に前後して前記上位装置に格納完了報告コマンドを送信する格納完了報告工程と、この格納完了報告工程の後前記キャッシュメモリから当該格納完了報告を行ったデータが削除されるまでの間に、前記各物理ディスクから当該データを読み出して前記キャッシュメモリに格納された対応するデータと比較するデータ比較工程と、このデータ比較工程で一致しない部分がある場合には当該一致しない部分を記憶した物理ディスクに障害が発生したと判定する障害判定工程とを備えたことを特徴とするアレイディスク制御方法。

【請求項２】　前記障害判定工程に続いて、当該障害と判定された物理ディスクに前記ディスクキャッシュに格納されたデータを再書き込みする再書き込み工程を備えたことを特徴とする請求項１記載のアレイディスク制御方法。

【請求項３】　前記障害判定工程に続いて、当該障害と判定された物理ディスクを縮退させて当該物理ディスクの使用を不可とする縮退工程を備えたことを特徴とする請求項１又は２記載のアレイディスク制御方法。

【請求項４】　前記格納完了報告工程に続いて、前記キャッシュメモリに格納されたデータのうち上位装置からのアクセス頻度の低いデータを選択する削除対象データ選択工程と、この削除対象データ選択工程で選択されたデータについて前記データ比較工程を実行させると共に前記障害判定工程によって障害がないと判定された場合に当該データを前記キャッシュメモリから削除する削除前比較工程とを備えたことを特徴とする請求項１記載のアレイディスク制御方法。

【請求項５】　上位装置から送信されたデータを一時的に記憶するキャッシュメモリと、前記上位装置から送信されたデータを物理ディスクの台数に応じて分割すると共に当該分割した分割データのパリティデータを生成するデータ分割部と、このデータ分割部によって生成された分割データ及びパリティデータを前記各物理ディスクにそれぞれ格納するデータ転送部とを備えると共に、

前記分割データ及びパリティデータとを各物理ディスクに格納した後、当該前記各物理ディスクから当該データを読み出して前記キャッシュメモリに格納された対応するデータと比較するデータ比較部と、このデータ比較工程で一致しない部分がある場合には当該一致しない部分を記憶した物理ディスクに障害が発生したと判定する障害判定部とを備えたことを特徴とするアレイディスク制御装置。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】本発明は、アレイディスク制御方法及び装置に係り、特に、データに冗長性を持たせて複数の物理ディスクにデータを分割して格納するアレイディスク制御方法及び装置に関する。

【０００２】

【従来の技術】従来より、上位装置から転送されたデータを分割し、さらに分割データのパリティデータを生成して、これらを別々の物理ディスクに格納することで、１つの物理ディスクに障害が発生した場合であってもデータを復旧可能な手法が用いられている。このようなデータ分割やパリティデータの生成については、例えばＲＡＩＤ３，５等にまとめられている。

【０００３】しかしながら、上記従来例では、物理ディスクに障害が発生し、データの読み出しはできるが、データの内容が不正である場合には、どの物理ディスクに障害が発生したのかを判定することができない。これに対し、特開平９－３０５３２８号公報記載の手法では、分割データからパリティデータを生成するのみならず、巡回冗長検査情報（ＣＲＣ）を付加している。そして、再生したデータのパリティが不正であった場合には、１台を障害ディスクとした場合の複数の組み合わせでデータを復元し、複数の復元データの内、ＣＲＣに基づいて正常であると判定された復元データを再生データとして上位装置に転送する。

【０００４】

【発明が解決しようとする課題】しかしながら、上記従来例では、データの読み出し時に不正データのチェックを行うため、データ書き込み後長時間データを読み取らなかった場合に、不正データが発生した原因の解明が困難で特定できず、すると、データ書き込み後長時間経過した場合には、同時に複数台の物理ディスクで障害が発生してしまうことも想定され、すると、データの復旧が困難となる可能性が高い、という不都合があった。

【０００５】

【発明の目的】本発明は、係る従来例の有する不都合を改善し、特に、不正データの発生及びディスク障害を早期に発見して長期間読み出されないデータであっても確実に保存することのできるアレイディスク制御方法及び装置を提供することを、その目的とする。

【０００６】

【課題を解決するための手段】そこで、本発明では、論理ディスクを構成する複数台の物理ディスクと、上位装置から送信されたデータを一時的に記憶するキャッシュ

メモリと、前記上位装置から送信されたデータを前記物理ディスクの台数に応じて分割すると共に当該分割した分割データのパリティデータを生成するデータ分割部と、このデータ分割部によって生成された分割データ及びパリティデータを前記各物理ディスクにそれぞれ格納するデータ転送部とを備えたディスクアレイ装置を使用して上位装置から送信されたデータを格納するディスクアレイ制御方法であって、分割データ及びパリティデータの前記各物理ディスクへの格納処理に前後して前記上位装置に格納完了報告コマンドを送信する格納完了報告工程と、この格納完了報告工程の後前記キャッシュメモリから当該格納完了報告を行ったデータが削除されるまでの間に、前記各物理ディスクから当該データを読み出して前記キャッシュメモリに格納された対応するデータと比較するデータ比較工程と、このデータ比較工程で一致しない部分がある場合には当該一致しない部分を記憶した物理ディスクに障害が発生したと判定する障害判定工程とを備えた、という構成を採っている。これにより前述した目的を達成しようとするものである。

【０００７】本発明では、上位装置から送信されたデータをまずキャッシュメモリに格納する。キャッシュメモリに当該データを格納すると、上位装置に格納完了報告コマンドを送信する。そして、キャッシュメモリに格納されたデータを分割すると共に、パリティデータを生成し、それぞれ複数の物理ディスクに格納する。上位装置への格納完了報告は、この分割データ及びパリティデータの物理ディスクへの格納が完了した時に行うようにしてもよい。

【０００８】続いて、上位装置から他のＩ／Ｏ要求がある場合にはそのＩ／Ｏ要求を処理する。そして、上記各物理ディスクに格納したデータと同一のデータをキャッシュメモリから削除する前までに、当該物理ディスクへのデータの格納が正常に完了したか否かの判定を行う。すなわち、データ比較工程では、各物理ディスクから当該データを読み出して前記キャッシュメモリに格納された対応するデータと比較する。そして、障害判定工程では、データ比較工程で一致しない部分がある場合には当該一致しない部分を記憶した物理ディスクに障害が発生したと判定する。このように、キャッシュメモリに格納されたデータと各物理ディスクに格納されたデータとを直接比較することにより、各物理ディスク毎に不正データの発生の有無を確認する。

【０００９】

【発明の実施の形態】以下、本発明の実施の形態を図面を参照して説明する。

【００１０】図１は本発明によるアレイディスク制御装置の実施形態の構成を示すブロック図である。図１に示すように、アレイディスク制御装置２は、上位装置１から送信されたデータを一時的に記憶するキャッシュメモリ６と、前記上位装置１から送信されたデータを前記物理ディスクの台数に応じて分割すると共に当該分割した分割データのパリティデータを生成するデータ分割部３と、このデータ分割部によって生成された分割データ及びパリティデータを前記各物理ディスクにそれぞれ格納するデータ転送部４とを備えている。データ分割部３は、データ転送部３に併設せず、データ転送部４の一機能としてもよい。

【００１１】アレイディスク制御装置２はさらに、分割データ及びパリティデータとを各物理ディスクに格納した後、当該前記各物理ディスクから当該データを読み出して前記キャッシュメモリに格納された対応するデータと比較するデータ比較部８と、このデータ比較部８で一致しない部分がある場合には当該一致しない部分を記憶した物理ディスクに障害が発生したと判定する障害判定部１０とを備えている。

【００１２】アレイディスク制御装置は、データ転送部４を介して、上位装置１と論理ディスク２０を構成する各物理ディスク２１，２２，２３と接続されている。ここでは、物理ディスクは、物理ディスク１から物理ディスクＮまでのＮ個設けられている。

【００１３】ディスク転送部４は、上位装置１とのデータの転送を行い、キャッシュメモリ６へのデータの読み取り／書き込み及び論理ディスク２０へのデータの転送を行う。キャッシュメモリは、物理ディスクよりアクセスが速い例えばＲＡＭ等の記憶媒体を使用する。

【００１４】データ比較部８は、上位装置から受け取った書き込みデータがキャッシュメモリ６に格納されている状態で、各物理ディスクに格納したデータを読み出してキャッシュメモリ６に格納されているデータと比較する。このとき、不一致部分があれば、正常に書き込まれていないこととなる。そして、正常に書き込まれなかった部分のデータが、どの物理ディスクに書き込まれたデータであるかは、データ分割部３による分割ルールに基づいて判明する。このため、比較したデータに不一致部分があれば、不一致を生じさせた不正データがどの物理ディスクに格納されていたかが判明する。

【００１５】従って、障害判定部１０は、データ比較部８で一致しない部分がある場合には当該一致しない部分を記憶した物理ディスクに障害が発生したと判定する。これにより、データの読み出しは可能ではあるが、正常に書き込みが行われないような障害が発生した物理ディスクを特定し、復旧処理を行うことができる。すると、長期に渡って読み出されないデータであっても、データの格納時に正確に格納されたか否かを確認するため、例えば複数台の物理ディスクが同時に障害を有しデータの復旧ができなくなる事態を防止することができる。また、データの格納時に正常に書き込みができなかった物理ディスクが特定されるため、当該物理ディスクの装置番号等を上位装置に書き込み障害情報として通知することで、不正データが生じた環境を維持したまま不正デー

タが生じた原因を解明することができ、従って、長期に渡った後に不正データが生じた場合と比較して、障害の原因解明が容易となる。

【００１６】復旧処理部１２は、障害判定部１０によって障害が発生した物理ディスクが特定されると、例えば、キャッシュメモリ６から不正データ部分を含む分割データを読み出して、再書き込み処理を１回又は数回繰り返し、再度キャッシュメモリに格納したデータと比較することで不正データが生じたか否かを判定する。このとき、当該不正データを生じさせた物理ディスクの異なるセクタへ書き込みを行うようにしても良い。このような再書き込みを行っても不正データが生じる場合には、当該不正データを生じさせた物理ディスクを縮退させる。縮退処理は、他のデータ全ての復元を行った後、当該物理ディスクを切り離す等の処理となる。

【００１７】図２は、上位装置（ホストプロセッサ）から書き込みデータとして送信されるデータの分割の例を示す説明図である。キャッシュメモリ６には、図２（Ａ）に示すようなホストプロセッサから受け取ったデータをそのまま格納する。一方論理ディスクに記録するデータは、ホストプロセッデータから受け取ったデータをデータ分割部３で分割して各物理ディスク１（符号２１）、物理ディスク２（符号２２）、物理ディスクＮ（符号２３）に記録する。Ｎ台の物理ディスクで構成される論理ディスク２０にデータを記録する場合、図２（Ｂ）に示すように、データをN-1個に分割して分割データ１（符号３１）、分割データ２（符号３２）から分割データＮ－２（符号３３）、分割データＮ－１（符号３４）を生成する。そして、分割したデータのパリティデータ３５を生成する。図２（Ｃ）に示すように、各物理ディスク１乃至Ｎには分割されたデータ１乃至Ｎ－１と、生成したパリティデータ３５が書き込まれる。データを読みとる場合、分割したデータ１乃至Ｎ－１とパリティデータ３５を全て読みとり、パリティチェックを行ってデータの妥当性を確認し、パリティデータ３５を除いたものをホストプロセッサ１に転送する。

【００１８】次に、図３を参照して図１に示したアレイディスク制御装置を使用してアレイディスクを制御する動作例を説明する。図３は本実施形態の動作例を示すフローチャートである。まず、データ転送部４が、上位装置（ホストプロセッサ）１から書き込みデータを受け取り（ステップＳ１）、キャッシュメモリ６へ当該データを格納する（ステップＳ２）。さらに、データ分割部３は、当該データを分割してパリティデータ３５を生成する（ステップＳ３）。データ転送部４は、この分割データ３１乃至３４及びパリティデータ３５を各物理ディスク２１乃至２３へ格納する（ステップＳ４）。続いて、データ転送部４は、ホストプロセッサ１へ格納完了報告を行う（ステップＳ５，格納完了報告工程）。具体的には、格納完了報告を示すコマンドを上位装置の入出力インタフェースへ送信する。この上位装置への格納完了報告は、実際に各物理ディスクにデータを格納した直後ではなく、ステップＳ２のキャッシュメモリ６への格納が完了した直後に行うようにしても良い。

【００１９】図３に示す例では、格納完了報告Ｓ５に続いて、すなわち、上位装置を切り離した後、物理ディスクに格納したデータとキャッシュメモリ６に格納されているデータとの比較を行う（データ比較工程Ｓ６乃至Ｓ８）。データ転送部４は、物理ディスクへ格納したデータを読み出す（ステップＳ７）。一方、データ比較部８は、キャッシュメモリ６から同一データであったデータを読み出し、パリティデータを生成する（ステップＳ８）。そして、データ比較部８は、この物理ディスクから読み出したデータとキャッシュメモリ６に格納されたデータとを比較する（ステップＳ９）。図３に示す例では、データ比較工程が、キャッシュメモリから読み出したデータのパリティデータを生成する工程を備えている。これにより、パリティデータを格納したデータの不正の有無を判定することができる。

【００２０】両データの比較の結果、一致していれば書き込みは正常に行われているため、データの書き込み処理を終了し、一方、両データに相違点がある場合には、当該一致しない部分を記憶した物理ディスクに障害が発生したと判定する（ステップＳ１０，障害判定工程）。そして、一方、両データに相違点がある場合には、復旧処理部１２を起動して復旧処理を行う（ステップＳ１１，復旧処理工程）。

【００２１】復旧処理工程Ｓ１１は、当該障害と判定された物理ディスクに前記ディスクキャッシュに格納されたデータを再書き込みする再書き込み工程や、当該障害と判定された物理ディスクを縮退させて当該物理ディスクの使用を不可とする縮退工程を備えると良い。

【００２２】図３に示す例では、上位装置１への格納完了報告Ｓ５が完了した直後に、物理ディスクへのデータの格納の直後にデータ比較工程を位置づけている。上位装置１からのデータの書き込み要求及び再生要求（Ｉ／Ｏ要求）が連続的であれば、図３に示すようにデータの書き込みの直後にデータ比較工程を実施することが好ましい。一方、上位装置のＩ／Ｏ要求が間欠的であれば、Ｉ／Ｏ要求を優先実行し、Ｉ／Ｏ要求の待機時間内を活用してデータ比較工程を実施するようにしても良い。この場合、本実施例ではキャッシュメモリ６に格納されたデータを利用するため、このキャッシュメモリ６に格納したデータを削除する前までには、データの比較を完了させなければならない。

【００２３】図４は、キャッシュメモリ６からデータを削除する直前までにデータ比較を行う処理例を示すフローチャートである。図４に示す例では、格納完了報告工程Ｓ５に続いて、前記キャッシュメモリに格納されたデータのうち上位装置からのアクセス頻度の低いデータを

選択する削除対象データ選択工程Ｓ２４と、この削除対象データ選択工程Ｓ２４で選択されたデータについて前記データ比較工程Ｓ９を実行させると共に前記障害判定工程Ｓ１０によって障害がないと判定された場合に当該データを前記キャッシュメモリから削除する削除前比較工程Ｓ２７とを備えている。

【００２４】具体的には、まず、上位装置からのＩ／Ｏ要求の有無を確認する（ステップＳ２１）。Ｉ／Ｏ要求が書き込みである場合、キャッシュメモリ６に十分な容量が無ければ実行できなくなるため、キャッシュの容量が十分であるか否かを確認する（ステップＳ２２）。キャッシュ容量が十分であれば、Ｉ／Ｏ処理を実行する（ステップＳ２３）。すなわち、データの書き込みが要求されれば、図３に示すステップＳ１乃至Ｓ５を実行し、データの読み出しを要求された場合には、キャッシュメモリ６に該当するデータがあれば当該データを上位装置に送信し、一方キャッシュメモリ６に該当するデータが無ければ物理ディスクからデータを読み出して復元の上、上位装置に転送する。データ読み出し要求時に、キャッシュメモリ６に当該データが格納されていた場合には、当該データのアクセス頻度を示す値を増加させるようにすると良い。

【００２５】上位装置からのＩ／Ｏ要求を待機している場合や、キャッシュ容量が不十分である場合には、まず、キャッシュメモリ６に格納されたデータの内、アクセス頻度の低いデータを検索する（ステップＳ２４）。このデータの検索・特定は、削除すべきデータの特定であるため、アクセス頻度の他、一定時間経過したデータを選択するなど、アレイディスクの分野で一般的な他の手法を採用することができる。

【００２６】削除候補となるデータが特定されると、当該データについて各ディスクに格納されたデータと比較する（ステップＳ２５）。すなわち、図３に示すステップＳ６乃至Ｓ９を実行する。そして、両データが一致していれば、当該データは正常に格納されているため、当該キャッシュメモリに格納されたキャッシュデータを削除する（ステップＳ２７）。キャッシュメモリに十分な空き容量がある場合には、削除可能であることを当該キャッシュメモリを管理する図示しないキャッシュメモリ管理部へ通知するようにしても良い。

【００２７】一方、データが一致しない場合には、図３に復旧処理Ｓ１１を実行する。

【００２８】上述したように本実施形態によると、キャッシュ上のデータとディスクに書き込まれたデータを比較することにより、上位装置からの書込データが複数の物理ディスクすべてに正しく書き込まれたか否かを判定することがき、すると、分割したデータの書き抜けを未然に防ぐ事ができるようになる。

【００２９】

【発明の効果】本発明は以上のように構成され機能するので、これによると、障害判定工程が、データ比較工程で一致しない部分がある場合には当該一致しない部分を記憶した物理ディスクに障害が発生したと判定するため、各物理ディスク毎に不正データの発生の有無をデータの書き込み直後に確認することができ、すると、物理ディスクの障害発生原因の解明が容易で且つ当該データを長期に渡って確実に保存することができ、しかも、キャッシュメモリに格納されたデータと物理ディスクに格納されたデータとを直接比較することで不正データを特定するため、不正データが格納されていた物理ディスクがどの物理ディスクであるかを即座に特定することができ、このため、障害復旧処理の選択肢が増え、例えば、複数回再書き込みを行った後にデータの不正が解消しない場合に当該物理ディスクについて縮退処理を行うなど、物理ディスクの障害の早期発見及び早期復旧を図ることができる、という従来にない優れたアレイディスク制御方法を提供することができる。

【図面の簡単な説明】

【図１】本発明の第一の実施形態の構成を示すブロック図である。

【図２】図１に示した各構成要素が扱うデータの構造を示す説明図であり、図２（Ａ）は上位装置から転送されるデータの構造を示す図で、図２（Ｂ）はディスク分割部によって生成されるデータの構造を示す図で、図２（Ｃ）は各物理ディスクに格納されるデータの構造を示す図である。

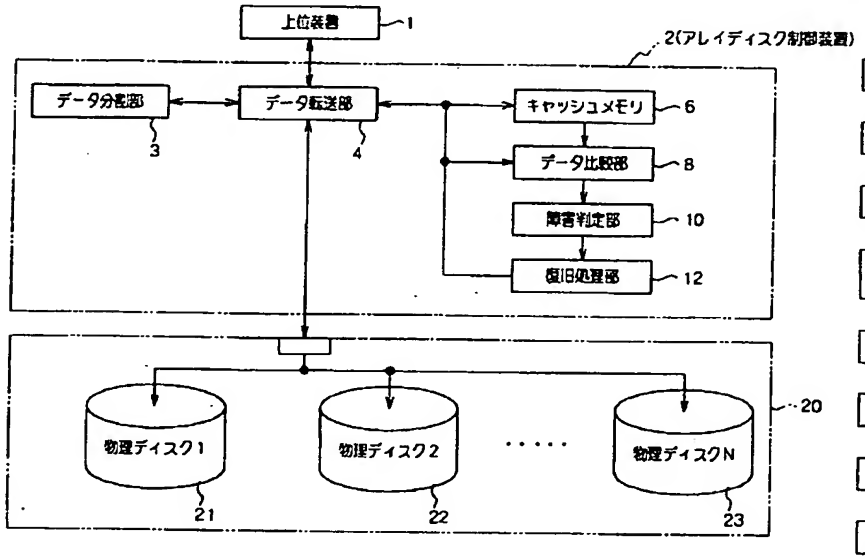
【図３】図１に示した構成でのデータ書き込み処理の一例を示すフローチャートである。

【図４】図１に示した構成での他のデータ書き込み処理の例を示すフローチャートである。

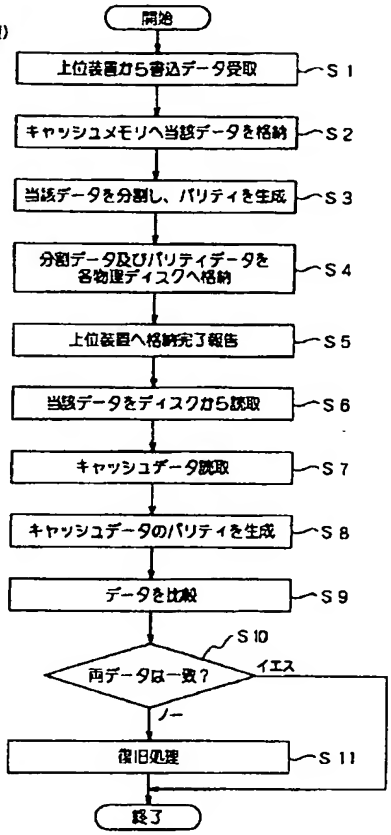
【符号の説明】

１　上位装置（ホストプロセッサ）
２　アレイディスク制御装置
３　データ分割部
４　データ転送部
６　キャッシュメモリ
８　データ比較部
１０　障害判定部
１２　復旧処理部
２０　論理ディスク
２１，２２，２３　論理ディスクを構成する物理ディスク

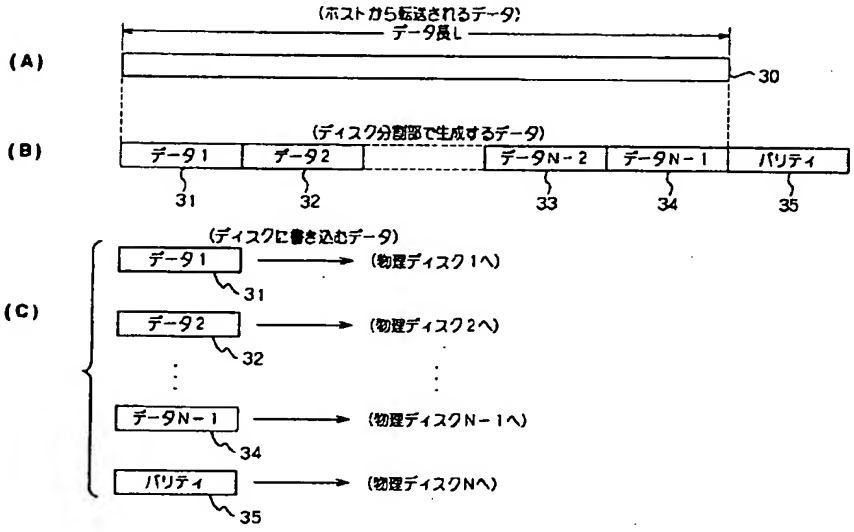
【図1】


上位装置
1
2(アレイディスク制御装置)
データ分割部
3
データ転送部
4
キャッシュメモリ
6
データ比較部
8
障害判定部
10
復旧処理部
12
20
物理ディスク1
21
物理ディスク2
22
物理ディスクN
23

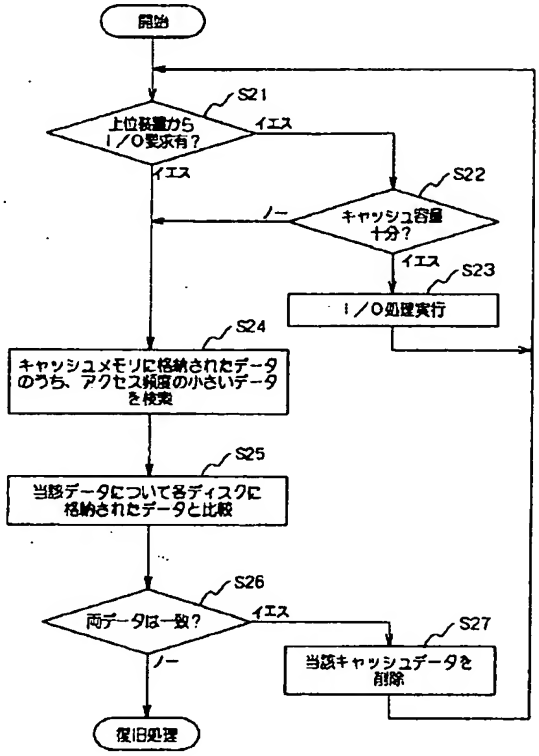
【図3】


開始
上位装置から書込データ受取
S1
キャッシュメモリへ当該データを格納
S2
当該データを分割し、パリティを生成
S3
分割データ及びパリティデータを各物理ディスクへ格納
S4
上位装置へ格納完了報告
S5
当該データをディスクから読取
S6
キャッシュデータ読取
S7
キャッシュデータのパリティを生成
S8
データを比較
S9
S10
両データは一致?
イエス
ノー
復旧処理
S11
終了

【図2】


(A)
(ホストから転送されるデータ)
データ長L
30
(B)
(ディスク分割部で生成するデータ)
データ1
データ2
データN-2
データN-1
パリティ
31
32
33
34
35
(C)
(ディスクに書き込むデータ)
データ1
(物理ディスク1へ)
31
データ2
(物理ディスク2へ)
32
データN-1
(物理ディスクN-1へ)
34
パリティ
(物理ディスクNへ)
35

【図4】



フロントページの続き

(51)Int.Cl.[illegible]	識別記号	FI	テーマコード(参考)
G11B [illegible]	501	G11B 20/18	501B
	522		522Z
	570		570Z
	572		572B
			[illegible]